University of Kent

Kent Academic Repository

Rizos, Georgios, Lawson, Jenna, Mitchell,, Simon L., Shah, Pranay, Wen, Xin, Banks-Leite, Cristina, Ewers, Robert and Schuller, Björn W (2024) *Propagating variational model uncertainty for bioacoustic call label smoothing.* Patterns (New York, N.Y.), 5 (3).

Downloaded from
https://kar.kent.ac.uk/105474/ The University of Kent's Academic Repository KAR

The version of record is available from
https://doi.org/10.1016/j.patter.2024.100932

This document version
Publisher pdf

DOI for this version

Licence for this version
CC BY-NC-ND (Attribution-NonCommercial-NoDerivatives)

Additional information

Versions of research works

Versions of Record
If this version is the version of record, it is the same as the published version available on the publisher's web site. Cite as the published version.

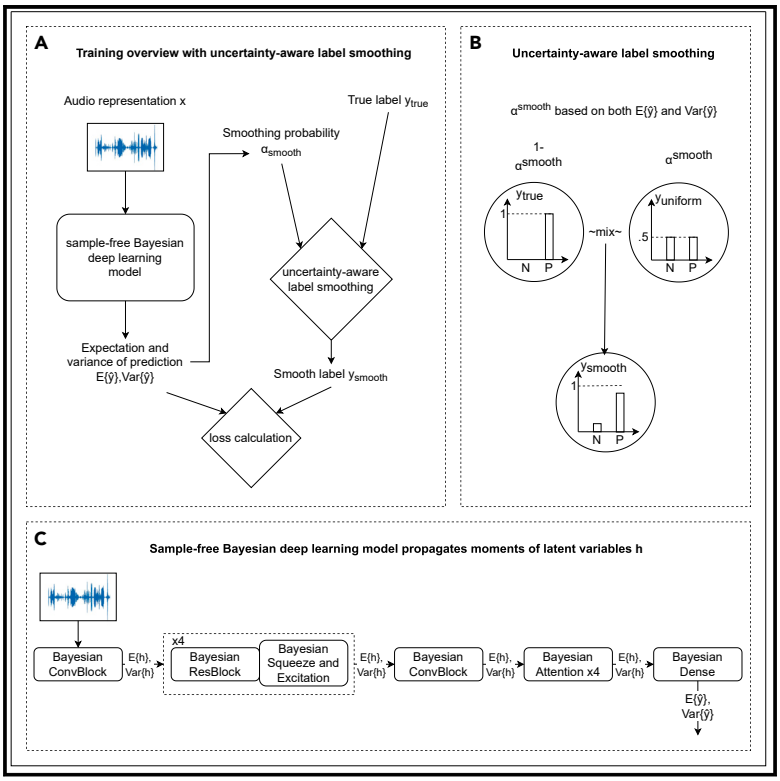
Author Accepted Manuscripts
If this document is identified as the Author Accepted Manuscript it is the version after peer review but before type setting, copy editing or publisher branding. Cite as Surname, Initial. (Year) 'Title of article'. To be published in ***Title of Journal*** , Volume and issue numbers [peer-reviewed accepted version]. Available at: DOI or URL (Accessed: date).

Enquiries
If you have questions about this document contact ResearchSupport@kent.ac.uk. Please include the URL of the record in KAR. If you believe that your, or a third party's rights have been compromised through this document please see our Take Down policy (available from https://www.kent.ac.uk/guides/kar-the-kent-academic-repository#policies).

Patterns

Propagating variational model uncertainty for bioacoustic call label smoothing

Graphical abstract



Highlights

- Sample-free, Bayesian attentive ResNet with squeeze and excitation
- Uncertainty-based, data-specific label smoothing
- Bioacoustic call detection on two datasets, one of which is introduced here
- Use of predictive uncertainty in label-smoothing parameterization

Authors

Georgios Rizos, Jenna Lawson, Simon Mitchell, ..., Cristina Banks-Leite, Robert Ewers, Björn W. Schuller

Correspondence

georgios.rizos12@imperial.ac.uk (G.R.), bjoern.schuller@imperial.ac.uk (B.W.S.)

In brief

In this article, sample-free Bayesian neural networks are applied to bioacoustic call detection in order to improve both predictive and calibration performance. The authors further explore the use of Bayesian predictive uncertainty to guide the training process to focus less on samples for which the model predicts higher uncertainty and show promising results on two animal call-detection datasets, one of which is introduced here.

Rizos et al., 2024, Patterns 5, 100932
March 8, 2024 © 2024 The Authors.
https://doi.org/10.1016/j.patter.2024.100932

CellPress

Patterns

CellPress
OPEN ACCESS

Article

Propagating variational model uncertainty for bioacoustic call label smoothing

Georgios Rizos,[1,5,*] Jenna Lawson,[2] Simon Mitchell,[3] Pranay Shah,[1] Xin Wen,[1] Cristina Banks-Leite,[2] Robert Ewers,[2] and Björn W. Schuller[1,4,*]

[1]GLAM – Group on Language, Audio, & Music, Department of Computing, Imperial College London, London SW7 2RH, UK
[2]Department of Life Sciences, Imperial College London, Ascot SL5 7PY, UK
[3]DICE – Durrell Institute of Conservation and Ecology, University of Kent, Canterbury CT2 7NR, UK
[4]EIHW – Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Augsburg, 86159 Bavaria, Germany
[5]Lead contact
*Correspondence: georgios.rizos12@imperial.ac.uk (G.R.), bjoern.schuller@imperial.ac.uk (B.W.S.)
https://doi.org/10.1016/j.patter.2024.100932

THE BIGGER PICTURE Uncertainty awareness in deep learning enables models to focus on learning from well-annotated data and to place less confidence on uncertain predictions. This has the potential to foster trust in algorithmic decision making and enhance policy making in applications pertaining to conservation using recordings made by on-site passive acoustic monitoring equipment. Such analyses can automate the annotation process and reduce human presence in the field.

SUMMARY

Along with propagating the input toward making a prediction, Bayesian neural networks also propagate uncertainty. This has the potential to guide the training process by rejecting predictions of low confidence, and recent variational Bayesian methods can do so without Monte Carlo sampling of weights. Here, we apply sample-free methods for wildlife call detection on recordings made via passive acoustic monitoring equipment in the animals' natural habitats. We further propose uncertainty-aware label smoothing, where the smoothing probability is dependent on sample-free predictive uncertainty, in order to downweigh data samples that should contribute less to the loss value. We introduce a bioacoustic dataset recorded in Malaysian Borneo, containing overlapping calls from 30 species. On that dataset, our proposed method achieves an absolute percentage improvement of around 1.5 points on area under the receiver operating characteristic (AUROC), 13 points in F1, and 19.5 points in expected calibration error (ECE) compared to the point-estimate network baseline averaged across all target classes.

INTRODUCTION

Effective wildlife monitoring can guide action to ameliorate the effects of the global biodiversity crisis but poses an enormous scalability challenge.[1,2] A potential solution for scalable bioacoustic data modeling[3] is offered by the combination of audio sensing infrastructure[4] and deep learning (DL), i.e., methods consisting of hierarchical stacking of linear processing layers and nonlinear pooling and activation operations. The monitoring of wildlife and environments using sound recorders—i.e., passive acoustic monitoring (PAM)[4]—allows for an automated, continuous monitoring solution that minimizes the duration of human presence in the field and, thus, the impact such presence can have on the behavior of the animals. Furthermore, the recordings no longer need to be limited to how much experts can reasonably listen, leading to great scalability both spatially and temporally. DL for bioacoustics offers the possibility of distilling the detection and categorization experience of ecology experts into a DL computational model. This can automate and expedite relevant labor, alleviating spurious annotation errors (as DL is known to be capable of doing[5]), such that the time of experts can be invested in a more fruitful manner. This scaled-up data enrichment can improve contributions to conservation- and ecology-related policy making.[6]

Many DL architectures that perform well in detecting specific signals in sound recordings—i.e., acoustic event detection (AED)—were originally designed for the visual classification domain.[7] For a recent example, residual networks (hence ResNets,[8] i.e., deep convolutional networks with residual connections every few layers for facilitating backwards propagation of the error signal for training) were shown to outperform the competition in a study on AED.[9] A ResNet similar to the winning

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

method from the aforementioned study was also shown to be the best performer specifically for bioacoustic call detection in an extensive comparative study[10] against a non-residual deep convolutional network,[9] shallower networks of around two or three (1D or 2D) convolutional layers commonly used for AED,[11–14] as well as a combination of convolutional and recurrent (i.e., designed for sequential data) layers previously used for the bioacoustic detection of Bornean gibbon calls.[15] The success of the winning model of Rizos et al.[10] was also due to the incorporation of attention methods, i.e., methods that entail the learning of weights that allow the model to focus on particular time frames[16,17] or convolutional filters.[18] Although both the former mechanism—attentive global sequence pooling—and the latter—squeeze and excitation (SE)—have been shown to be contribute to improvements in the acoustic domain as well,[12,19] including to the previously mentioned improved variant of ResNet for call detection,[10] they have not necessarily been adopted in later acoustic ResNet-based call-detection studies.[20,21] A recent alternate approach is BirdNET,[20] an application of the Wide ResNet[22] model (a variant of ResNet using larger filter numbers) on a large composite dataset on detection of calls from 984 bird species that achieves competitive results compared to other methods that have been tested on similar datasets with much fewer species. The pretrained BirdNet model has since been extensively applied on various datasets.[23] Finally, although in this study we focus on the task of call detection, related tasks include cross-[24] or within-species[25] call type classification and individual identification.[26] Such liberally selected applications exist across a wide range in animal species, e.g., on primates,[10,14,15,27] whales,[24] and birds.[28,29]

It is important, however, that the predictions made by the DL model are understood and trusted. Unfortunately, during this near-decade of DL advancement, a fixation by the DL community toward deeper and more complicated architectures, as well as on traditional prediction performance evaluation measures, has led to an insidious DL model behavior manifesting overconfident predictions,[30] i.e., predictions made at a probability nearing 1, regardless of whether they are correct or not. Downstream software modules or policy makers making catastrophic decisions due to these overconfidently predicted misclassifications can foster deep mistrust in DL,[30–32] something that has also been noted with respect to bioacoustics.[3] However, early prediction calibration fixes[30] are based on learning a transformation of the model outputs that requires the existence of a validation set of labels, something that cannot be safely assumed in general. Another approach is label smoothing,[33] a regularization method that has also been used with the intention of improving calibration.[34] A smoothing probability hyperparameter, selected *a priori*, allows us to treat a ground-truth label annotation as noisy instead of binary (e.g., probability of 0.9 of a call being present, instead of a fully confident 1). Although it was originally proposed as a means to improve predictive performance,[33] its success in that regard[35,36] has also been inconsistent, as it has in other cases been shown to deteriorate it,[34,37] without necessarily improving calibration.[34] Label smoothing has also shown promise[21] in some cases on a call-detection study; however, no evaluation of calibration was made.

A means of designing DL models with the ability to accompany their standard predictive output with a measure of uncertainty is Bayesian inference. Predictive uncertainty is a signal that the input sample may have potentially been mislabeled.[38] Bayesian neural networks (BNNs)[39] have been shown to naturally offer better calibrated outputs as well as regularization compared to non-Bayesian, point-estimate versions of the same underlying architectures[40] (see related surveys on in-depth discussion for why this happens,[41] as well as lists of domain applications[42]). This is due to the predictive uncertainty, which describes a distribution from which less overconfident predictions can be sampled. BNNs employ distributional weight parameters, of which the posterior distributions are calculated via Bayes' rule and dependent on the observed training set and a prior distribution assumption.[39] Since, however, the integration for these posteriors is intractable (due to containing high dimensionality factors, see Blei et al. and Zhang et al.[43,44]), marginalizing the weights in order to get the statistical distribution of the outputs is often approximated via Monte Carlo (MC) sampling. As the uncertainty of stochastic parameters informs the output of each layer, and, hence, the input of each subsequent layer, we can understand the uncertainty information being propagated through the entire network until the final layer calculates the output (or epistemic) uncertainty. Using MC-based approximation to calculate it, one has to use K MC samples, something that increases the computational load by K. MC-based approaches comprise Bayes by backprop[45] and MC dropout,[46] and have been applied on a wide range of data domains, including audio.[13,47]

Uncertainty propagation in an MC sample-free manner can be performed by the approximation of the first two moments (i.e., expectation and variance) of the layer output pre-activations by leveraging the central limit theorem (CLT). This approach was used first for fast dropout,[48] where it allowed for sampling from the much fewer pre-activations instead of the layer weights, and later in the context of BNNs.[49] Later sample-free BNNs use closed-form, uncertainty-propagating, nonlinear activation functions[50–54] and eschew the need for sampling even from pre-activations. Apart from avoiding costly weight sampling, this approach is also not subject to the stochasticity of MC-based approaches. This has been hypothesized to be the reason behind their improved performance compared to MC-based methods in prediction and calibration performance.[53–55] Propagation of more than two moments has been shown to be beneficial, e.g., in resisting adversarial attacks (i.e., target distortion of test data such that the output is misclassified[56]) but also requires sampling for cubature,[57] or unscented[55] and particle[58] filtering. Such models have been applied on computer vision tasks such as image classification and segmentation, on data ranging from standard benchmarks[54] such as CIFAR,[59] to medical and radar images,[55] but never to audio, and, specifically, to bioacoustic call detection.

That being said, many recent moment-propagating BNN studies constitute Bayesian treatments of DL models with simple mechanisms, such as dense[52,53,58] and convolutional[51,55] layers interweaved by nonlinear activation functions,[50,52,53,55] even in non-Bayesian uncertainty propagation.[60] Although a sample-free Bayesian version of a dense layer-based network with ResNet-like skip connections has been proposed in Wu et al.,[53] less consideration has been given on doing the same for more advanced concepts such as convolutional ResNets, SE, and attention. Furthermore, even though the sample-free Bayesian approach has been shown to be superior to MC-based

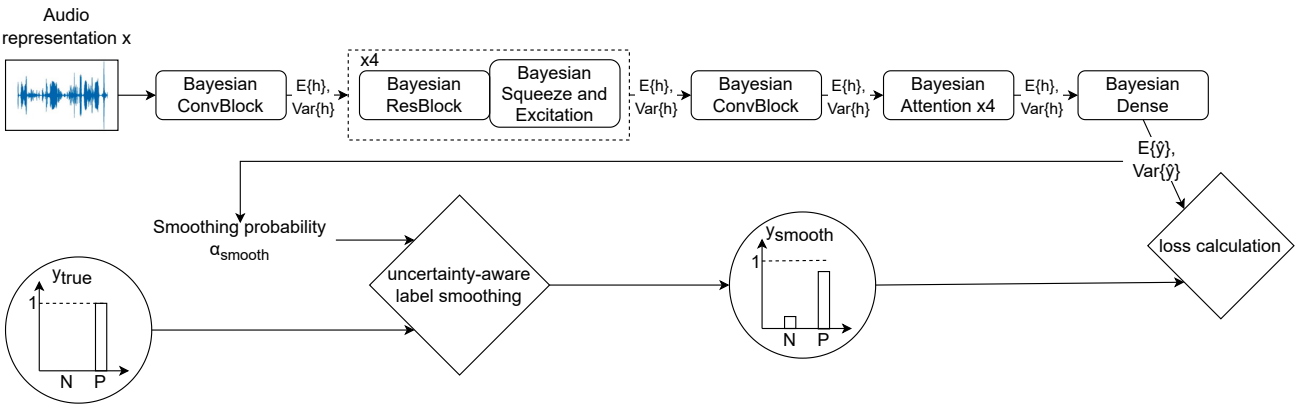


Figure 1. This is an abstraction of the sample-free, moment-propagating variational Bayesian SE-ResNet model with multi-head attention we use as a basis throughout this study

The point-estimate version follows the same architecture, but each layer, block, and nonlinearity does not use variational learning for inference or make affordances for propagating uncertainty. The outputs of the Bayesian SE-ResNet are used to parameterize a label-smoothing operation, and the loss calculation is performed using the smooth label.

BNNs,[53–55] only the latter approach has been used in bioacoustics[13] (and, in fact, on a shallower three-layer network of the kind that has been shown to underperform compared to deeper ResNets[10]).

Despite the demonstrated promise of sample-free BNNs, there does not exist an explicit utilization of sample-free predictive uncertainty as a signal for data-specific regularization during training. We believe that such an explicit usage can guide the model to not place as much weight on the learning of data that it calculates as noisily annotated, something that can potentially improve both predictive and calibration performance. We also believe that this is a very timely topic for investigation in the domain of bioacoustic call detection, a domain where the need for calibration of model output probabilities (along with traditional accuracy-based performance evaluation) has been repeatedly suggested.[3,61] This is especially important as probabilistic, instead of categorical, outputs are considered to be more informative for downstream decision making.[62]

The contributions we make in this article are summarized as follows.

(1) We perform the first exploration of sample-free, uncertainty propagating, variational Bayesian DL on bioacoustic call detection in order to exploit the regularization and the better calibration that such models exhibit. Specifically, we provide a sample-free Bayesian treatment of a complex DL architecture that has excelled in the call-detection task.[10] It propagates activation expectations and variances through mechanisms such as global attention pooling and SE blocks. To our knowledge, this is the first time a moment-propagating version of the SE mechanism has been proposed and evaluated, although MC-based Bayesian methods have done so before.[63] We further consider two variants of the underlying model concerning the type of local pooling: one using the known[55,64] moment-propagating version of max-pooling, and a moment-propagating version that we first use of an attention-pooling method inspired by recent studies.[65,66] Our results indicate that opting for a sample-free Bayesian DL method is indeed the most promising approach as it outperforms the corresponding point-estimate baseline in most cases.

(2) We propose a regularization method that explicitly uses the propagated predictive uncertainty of a sample-free BNN model as a signal for adaptive label smoothing that is specific to each data sample. The rationale is that the importance of highly uncertain samples could be attenuated in the loss calculation. An overview of the whole approach is depicted in Figure 1. This approach achieves generally higher predictive and calibration performance compared to our other baselines when the underlying model uses maximum local pooling. In the case of attention local pooling, the comparison is less conclusive, as the best performer is either a variant in which the same smoothing probability is used for all samples in a batch (hence, data-sample agnostic) or no label smoothing at all. Our results indicate, however, that deterministic, moment-propagating BNNs—including our proposed method—exhibit high calibration performance also in bioacoustic call detection compared to point-estimate networks.

(3) Our methodology is evaluated on challenging, real-world, "in-the-wild" datasets, as literally is the case in bioacoustics for wildlife PAM. The recordings may contain multiple background sounds other than the target calls. We obtain the best reported results on a spider monkey call-detection dataset previously used in Rizos et al.,[10] and further introduce a new dataset with annotations for 30 distinct species (29 bird species, and Bornean gibbons) with potentially overlapping calls. The latter, which we call the SAFE Project[67] Multi-Species Multi-Task (SAFE-MSMT) dataset, is available at Zenodo: https://doi.org/10.5281/zenodo.7740620.

RESULTS

The common type of task between our two animal call-detection datasets is binary classification (i.e., positive class when one or

Table 1. SE-ResNet with multiple-head attention implementation

Model operation	Shape
Log-Mel spectrogram	(300, 128)
(ConvBlock @ 64, ReLU) & Pool	(150, 64, 64)
(SEBlock @ 64, ReLU) × 2 & Pool	(75, 32, 64)
(SEBlock @ 128, ReLU) × 2 & Pool	(37, 16, 128)
(SEBlock @ 256, ReLU) × 3 & Pool	(18. 8, 256)
(SEBlock @ 512, ReLU) × 2 & Pool	(9, 4, 512)
(ConvBlock @ 1024, ReLU)	(9, 4, 1024)
Reshape embedding	(9, 4096)
4-head attention-based pooling	(4096 × 4)
Dense layer per task	(1) × tasks

The sample-free variational versions share the same architecture, albeit by propagating moments throughout.

more calls of a particular type are found in a recorded clip, negative class otherwise). The Osa Peninsula Spider Monkey Whinny (OSA-SMW) dataset was first introduced and described in Rizos et al.,[10] and a single binary call-detection task is defined on it, where the focus is specifically the whinny call of Geoffroy's spider monkey (*Ateles geoffroyi*). We first introduce here the SAFE-MSMT dataset, of which the description and preprocessing details can be found in supplemental experimental procedures (sub-section "SAFE-MSMT Dataset"). We consider the detection of calls for each species identified within the dataset as a separate binary task and have identified 30 species such that, for all tasks, there are positive examples for each class in all of the training, development, and testing sets. It is possible that there are zero, one, or more species' calls audible per audio clip, which constitutes a multi-label classification problem. We approach this via a multi-task framework where each independent task is binary classification. This is realized by having one prediction layer per task, responsible for predicting the probability of the presence of a corresponding species call.

For evaluating our experiments, we opted to report the non-interpolated area under the precision-recall (AU-PR) curve of the positive class, and the area under the receiver operating characteristic (AU-ROC) curve as prediction performance measures that average over all possible probability thresholds for classification. Test performance is measured using the model that achieved best validation performance according to AU-PR, which is a stricter measure in class-imbalanced cases where the positive class is a minority, as AU-ROC is known to inflate due to the abundance of true negatives. We also report the unweighted average of F1 of the positive and negative classes at a probability threshold of 0.5 (F1) as well as the expected calibration error (ECE) for measuring calibration quality, as suggested by Guo et al.,[30] with 10 probability buckets. In order to provide a summary performance profile for the 30-task SAFE-MSMT dataset, we report here the weighted average of the per-task performance measures, where each weight is proportional to the number of positive instances per task. Even so, this is a quite austere evaluation as, for some species, there are only a handful of positive samples (as low as four), which heavily restricts the predictive potential of supervised-learning-based approaches.

As the baseline in our comparisons, we used a variation of a modern, complex DL model that was the best-performing method in a comparative study on the bioacoustics domain.[10] It combines a ResNet architecture, SE blocks, and multi-head global attentive pooling of sequential embeddings, and an output dense layer per binary task, for a total depth of 21 layers (instead of 28 in Rizos et al.[10]); hence, base SE-ResNet. A summary of its architecture, including parameter values and tensor shapes, can be found in Table 1, and more details are given in section "description of multi-attentive SE-ResNet." It is designed to process log-Mel spectrograms as input, i.e., two-dimensional audio representations.

We compare the performance of the base SE-ResNet with those of the (1) uncertainty propagating, variational Bayesian version developed for this article in section "crafting a competitive Bayesian SE-ResNet baseline") variant with the addition of our sample-free, uncertainty-aware label-smoothing technique in section "benefits of uncertainty usage in label smoothing," a pictorial overview of which can be seen in Figure 1. In an effort to show whether our proposed approach is robust to variations in the base architecture, we identify the *local* pooling operation as a point of interest. This is due to it being less explored in cited related literature on sample-free Bayesian DL,[52–55] where only the max-pooling (max-pool) equivalent operation is considered. We further consider an attentive pooling (att-pool) operation that is similar to the recent eMPool[66] and local importance pooling[65] operations. Our att-pool employs an additional dense layer and a softmax nonlinear activation that learn a weighted average of the activations to be pooled. More details on the implementation of core mechanisms, the considerations made toward a Bayesian treatment, and technical propositions can be found in section "experimental procedures," and full technical details in the supplemental experimental procedures. We summarize in Table 2 the predictive and calibration performance measure results that arose from our comparative analysis on animal call detection, which includes sample-free BNNs (for a higher granularity report of certain endangered species from SAFE-MSMT; see Table S1). In all cases, we performed eight trials for which we report mean and standard deviation.

SE-ResNet is a competitive point-estimate baseline

Although our goal is to show the benefits of sample-free Bayesian DL (with and without uncertainty-aware label smoothing) on bioacoustic call detection, we nevertheless performed one point-estimate neural architecture comparison, with a Wide ResNet[22] that was used in a bird call classification study (BirdNET[20]). We made our own implementation of the architecture, and train it from scratch on the datasets we include in our study using the same setup as our own methods. This is done in the interest of a fair comparison and because the pretrained BirdNET is trained to predict neither all the bird species in our SAFE-MSMT dataset nor spider monkey whinnies from OSA-SMW. The results of the comparison with our SE-ResNet (both the maximum and attention-pooling versions) are summarized in Table 3. We continue, thus, with sample-free, Bayesian treatments of only SE-ResNet in the following.

Crafting a competitive Bayesian SE-ResNet baseline

As a first step toward a more uncertainty-aware approach, we modify base SE-ResNet such that it becomes a variational Bayesian, uncertainty-propagating version of itself. Linear operators such as dense and convolutional neural layers are replaced

Table 2. Comparative study on two datasets between point-estimate neural networks and their sample-free Bayesian DL versions with and without uncertainty-aware label smoothing

SAFE Project Multi-Species Multi-Task					
	SE-ResNet	W-AU-PR ↑	W-AU-ROC ↑	W-F1 ↑	W-ECE ↓
max-pool	base	21.16 ± 2.16	78.45 ± 2.35	36.31 ± 11.94	35.86 ± 11.31
	variational	22.44 ∗ ± 2.00	79.16 ± 1.75	46.68 ± 4.33	22.63 ± 4.77
	smooth	22.25 ± 1.11	79.83 ± 2.89	52.43 ∗ ± 3.35	17.00 ± 7.19
	ua-smooth	20.76 ± 2.57	80.05 ∗ ± 2.81	49.61 ± 3.71	16.21 ∗ ± 3.19
att-pool	base	16.01 ± 2.25	72.15 ± 3.19	39.51 ± 13.37	29.36 ± 12.74
	variational	20.38 ∗ ± 2.70	77.97 ∗ ± 2.09	47.96 ∗ ± 3.61	21.63 ± 4.67
	smooth	15.53 ± 3.33	65.35 ± 7.90	47.86 ± 7.85	18.96 ∗ ± 13.66
	ua-smooth	16.94 ± 2.11	69.82 ± 4.44	38.75 ± 10.83	31.81 ± 11.72
Osa Peninsula Spider Monkey Whinny					
	SE-ResNet	AU-PR ↑	AU-ROC ↑	F1 ↑	ECE ↓
max-pool	base	81.81 ± 2.46	97.01 ± 0.79	82.95 ± 4.44	3.51 ± 1.30
	variational	82.74 ± 1.14	97.14 ± 0.34	80.31 ± 3.14	4.56 ± 1.16
	smooth	82.55 ± 1.60	97.26 ± 0.43	82.79 ± 4.17	3.66 ± 1.35
	ua-smooth	83.79 ∗ ± 2.42	97.47 ∗ ± 0.38	83.40 ∗ ± 3.22	3.46 ∗ ± 1.21
att-pool	base	84.81 ± 0.93	97.41 ± 0.34	84.38 ± 3.79	3.32 ∗ ± 1.63
	variational	84.82 ± 1.94	97.28 ± 0.55	78.74 ± 6.82	5.18 ± 3.36
	smooth	85.83 ∗ ± 0.60	97.47 ∗ ± 0.37	84.89 ∗ ± 5.85	3.53 ± 2.49
	ua-smooth	82.24 ± 5.42	96.68 ± 1.54	81.32 ± 4.90	3.99 ± 1.72

The proposed ua-smooth method distinguishes itself in the case where max-pool is used by the SE-ResNet. In case att-pool is used, the highest performer is either variational for the SAFE-MSMT dataset or smooth for OSA-SMW. The choice of max-pool works better for SAFE-MSMT, whereas att-pool works better for OSA-SMW, thus the use of label smoothing and whether it is uncertainty aware or not should be made depending on dataset. We denote by asterisks the best value (%) for each performance measure per dataset and per pooling type choice in order to more easily track the comparisons among methods based on the same backbone architecture. We further denote by italics the highest value per dataset, regardless of pooling choice.

with locally reparameterized versions, as described respectively in Kingma et al.[49] and Shridhar et al.[51] The first two moments of the outputs are given in closed form and are linearly dependent on the, also stochastic, respective layer inputs and weights, yet independent among themselves. The stochastic layer outputs are transformed by nonlinear activation functions such as ReLU and sigmoid, where the first two moments of the activations are approximated as previously described elsewhere.[48,50,52,60] Regarding max-pooling of such normally distributed variables, the authors of several studies[55,64] independently proposed co-pooling of the two moments, i.e., propagating only the moments of the random variable with the highest expected value. As for attention pooling, the weighted sum of normally distributed variables is well known, and we learn the probabilistic weights using attention. This way, information on the first two moments of all pooled variables is propagated.

Table 3. Comparison of point-estimate neural network baselines

SAFE Project Multi-Species Multi-Task				
Model	W-AU-PR ↑	W-AU-ROC ↑	W-F1 ↑	W-ECE ↓
SE-ResNet (max)	21.16 ∗ ± 2.16	78.45 ∗ ± 2.35	36.31 ± 11.94	35.86 ± 11.31
SE-ResNet (att)	16.01 ± 2.25	72.15 ± 3.19	39.51 ± 13.37	29.36 ± 12.74
Wide ResNet	19.75 ± 2.00	77.82 ± 1.88	52.51 ± 3.99	12.32 ∗ ± 5.01
Osa Peninsula Spider Monkey Whinny				
Model	AU-PR ↑	AU-ROC ↑	F1 ↑	ECE ↓
SE-ResNet (max)	81.81 ± 2.46	97.01 ± 0.79	82.95 ± 4.44	3.51 ± 1.30
SE-ResNet (att)	84.81 ∗ ± 0.93	97.41 ∗ ± 0.34	84.38 ∗ ± 3.79	3.32 ∗ ± 1.63
Wide ResNet	74.79 ± 2.15	95.62 ± 0.57	76.22 ± 5.97	5.30 ± 2.69

The comparison is among our implementations of a Wide ResNet previously used for bird classification,[20] an SE-ResNet previously used on the OSA-SMW dataset,[10] and a variation of the latter using attention local pooling. Although the Wide ResNet achieves the best performance in W-F1 and W-ECE for the SAFE-MSMT dataset, it is surpassed by SE-ResNet with max-pooling in the other two measures. Furthermore, it is surpassed by both SE-ResNet variants in all measures in the OSA-SMW dataset. We denote by asterisks the best value (%) for each performance measure per dataset.

Table 4. Training and prediction batch execution times in milliseconds for a batch size of eight on the SAFE-MSMT dataset

	SE-ResNet	Training time	Prediction time
max-pool	base	112	32
	variational	706	166
	smooth	714	166
	ua-smooth	713	166
att-pool	base base	134	40
	variational	753	176
	smooth	763	176
	ua-smooth	760	176

We measure both training time (including backpropagation) and prediction time. Regarding training times, the Bayesian methods are $\sim$ 6.3 and $\sim$ 5.6 times slower compared to the point-estimate baseline in the max-pool and att-pool cases, respectively. Regarding prediction times, the factors are, instead, $\sim$ 5.2 and $\sim$ 4.4.

In the results shown in Table 2, the Bayesian, uncertainty-propagating version of SE-ResNet with max co-pooling (variational max-pool) exhibits a slightly higher performance than base max-pool in terms of AU-PR, and AU-ROC for the OSA-SMW dataset, and an improvement across all measures for SAFE-MSMT, including the highest AU-PR among all max-pool-based methods.

As for the attention-pooling variant (variational att-pool), we observe a higher performance compared to the point-estimate baseline in all measures for SAFE-MSMT but at the same or lower performance in all measures for OSA-SMW.

Benefits of uncertainty usage in label smoothing

Label smoothing[33] in loss calculation is the use of a label distribution that is an interpolation between the true distribution, as given by the annotators, and the uniform distribution. In the binary classification task, the latter corresponds to 0.5 probability for both the negative and the positive classes:

$$y_{i,c}^{smooth} = \alpha y_{i,c}^{uniform} + (1 - \alpha) y_{i,c}^{true}, \quad \text{(Equation 1)}$$

where $y_{i,c}$ refers to the label probability that class c is correct for data sample i, and α denotes the smoothing probability hyperparameter. The latter quantifies the degree to which we want the model to *not* overexert in trying to learn to classify that particular sample as per the ground truth $y_{i,c}^{true}$.

Here, we propose a solution for data-specific label smoothing that is dependent on the uncertainty propagated throughout a BNN model, and is also MC sample free. A description of the means by which we define such an uncertainty-aware smoothing probability α_i for sample i, is found in section "experimental procedures," and a schematic overview is depicted in Figure 1.

As seen in the results in Table 2, our uncertainty-aware label-smoothing method (ua-smooth) used on the BNN described in section "crafting a competitive Bayesian SE-ResNet baseline" outperforms the max-pool-based variational method in terms of all measures except for W-AU-PR on the SAFE-MSMT dataset. In the att-pool case, we do not observe a similar behavior, as the only improvement is on F1 and ECE in the OSA-SMW dataset. In the max-pool case, the ua-smooth method also achieves better performance compared to the baseline in all cases.

Smoothing should be specific to data samples

How can we be sure, then, that the propagated model uncertainty contains information about which samples should use higher smoothing probabilities and that it is not simply a case of label smoothing being beneficial in general?

To answer this question, we perform one more series of experiments, with a label-smoothing variant (hence, smooth) that keeps the smoothing probability fixed across the training batch. Specifically, we calculate for every batch the average of the uncertainty-aware smoothing probabilities as per our proposed ua-smooth method and apply that to all batch samples instead. This is not a hyperparameter-based, fixed-value label smoothing, as is commonly used, since it benefits from the uncertainty quantification provided by the BNN, the values of which change per training step as the model learns to model the training data, and it tracks the average value of the uncertainty-aware smoothing probability, thus allowing for a stricter comparison with the ua-smooth method, which we propose as the better means of performing uncertainty-aware label smoothing using a BNN.

We observe from Table 2 that, in the max-pool case, ua-smooth always outperforms smooth for all measures on OSA-SMW, whereas on SAFE-MSMT this holds only for W-AU-ROC and W-ECE. In the att-pool case, it is instead smooth that outperforms ua-smooth for all measures on OSA-SMW, whereas the comparison is also inconclusive on SAFE-MSMT, with ua-smooth performing better in terms of W-AU-PR and W-AU-ROC only.

Sample-free BNN outputs are calibrated

A recommendation on which Bayesian approach to use agnostically is not easy to make, although, on the SAFE-MSMT dataset, the calibration performance of the point-estimate baselines are worse compared to all corresponding Bayesian versions. We do not observe the same behavior on the OSA-SMW dataset, although, in the max-pool cases on both datasets, it is our proposed ua-smooth method that achieves the best ECE performance among the corresponding competing methods.

Locally pooling normal random variables

Apart from max-pooling, we have also showcased the efficacy of a BNN approach based on newer, more elaborate local pooling methods.[65,66] Although we see that, on the OSA-SMW dataset, attention pooling brings a clear improvement on all performance measures over the use of max-pooling, on SAFE-MSMT the behavior is reversed; i.e., max-pooling is overall the best-performing local pooling operation. We find that our proposed ua-smooth method manages to achieve best performance compared to corresponding competing methods for the max-pooling case, but not for attention pooling, where either the naive smooth method works best on OSA-SMW or the sample-free BNN without any label smoothing in SAFE-MSMT.

Execution times

We further perform a wall-clock execution time measurement for all the competing methods on a machine equipped with an Nvidia GeForce GTX 1080 Ti graphics processing unit (GPU) with 11 Gb of memory. The results are summarized in Table 4. The increase in execution times for the sample-free Bayesian methods is well known in relevant literature.[53–55]

The time per epoch of training is dependent on dataset size. For example, for SAFE-MSMT and using the max-pooling variants, an epoch of training requires $\sim$21 s and $\sim$140 s for point-estimate and Bayesian versions, respectively, whereas for OSA-SMW it is $\sim$88 s and $\sim$560 s. For SAFE-MSMT and using max-pooling, training requires around 20 min and 3 h for point-estimate and Bayesian versions, respectively. For OSA-SMW, the training times are, correspondingly, 1.5 h and 10 h. The higher overall training times for Bayesian methods can be explained by the fact that they require more epochs as they generally reach better parameter set optima.

DISCUSSION

We now discuss (1) the insights extracted from our experiments regarding our proposed methodology in section "propagated uncertainty should be explicitly used"; (2) relations to similar methods and means by which our method should engender a re-evaluation thereof in section "rethinking label smoothing"; and (3) potential extensions, criticisms, and opportunities in sections "should we focus on the easy data then?" to "conclusions and future work."

Propagated uncertainty should be explicitly used

Propagated predictive uncertainty, as per our variational variant of SE-ResNet, affects loss value calculation as it describes a predictive distribution from which multiple prediction instances can be sampled. This leads to an expected loss value calculation that is based on softer, less overconfident prediction outputs compared to a loss value based on point-estimate predictions; the utilization of epistemic uncertainty involving all potential output samples has been cited as a major regularizing strength of BNNs.[41]

In addition to the point-estimate base, we have designed the sample-free variational method to be a more advanced baseline, to more strictly compete with our proposed uncertainty-based label-smoothing method.

That being said, by means of an insight from our experiments with the moment-propagating "flavor" of BNNs, i.e., the variational Bayesian SE-ResNet, we observed promising (e.g., overall improvement on the OSA-SMW dataset) yet inconclusive results. As such, we recommend that the Bayesian property, as well as the type of uncertainty-aware label smoothing, should be considered to be types of hyperparameters, not to be employed agnostically but only after experimental validation on the task under examination, including consideration of the relevant performance measures thereof.

However, the Bayesian formulations offer us another highly informative signal, something exclusive to them and unavailable to the baseline: the value itself of predictive variance, i.e., a proxy of epistemic uncertainty. There is a more explicit manner of utilizing it, which can, and indeed should, be used in the loss calculation, as, in our experiments, the ua-smooth method performs better than the corresponding variational in most performance measures in the case of models using max-pooling.

Usually, predictive uncertainty is used in downstream tasks, e.g., as a signal for data acquisition in active learning,[52,68] or toward the design of uncertainty-aware (e.g., risk-averse) reinforcement learning agents.[69] Inversely, we believe that uncertainty should be used as a signal that guides learning in the self-same task, and by the self-same model that is undergoing training; as per our experiments, not doing so may lead to missing the opportunity given by the usage of a BNN and is also disregarding one-half of the BNN output. The sample-free manner of uncertainty offers a more elegant and less stochastic means of doing so compared to MC-based methods.

More than that, our experiments with the batch-wide fixed smoothing method (smooth) indicate that a higher degree of label smoothing can be beneficial to data samples for which the BNN is less confident in modeling, and that, thus, the ua-smooth variant is preferrable. That being said, for the OSA-SMW dataset in the att-pool case, it seems that smooth performs better than the other corresponding methods, indicating that Bayesian regularization may be beneficial for that dataset in any shape or form, most probably due to the positive class sample scarcity in all binary classification tasks of this dataset.

Rethinking label smoothing

That being said, label smoothing has been considered as one of the reasons for the high performance achieved by the student model in knowledge distillation[70]; i.e., a learning framework involving a student model learning from the predictions of a teacher model that is itself trained with the true labels. Knowledge distillation utilizes the smooth prediction probabilities output by the teacher model in place of ground-truth labels. These output distributions are smoother, i.e., closer to the uniform, for data samples that the teacher model finds difficult to model, thus constituting data-specific smoothing. Moving away from the two-step, teacher-student framework (that is focused on model compression), in this study, we have shown the usefulness of a means for smoothing that requires no more than a single model, a single training process, and is also MC sample free. As indicated by our experiments, we believe that the underlying conception of label smoothing is still promising, with the caveat that they need to be made in an adaptive, intelligent, and data-specific manner; *a fortiori* in the uncertainty-propagating BNN context, where a guiding signal is provided by design.

The study that is closer conceptually to our own, in terms of attempting to improve accuracy and calibration, is the one performed in Seo et al.,[71] in which the authors use the MC-based BNN approach proposed in Gal and Ghahramani[46] called MC dropout and focus on image classification. They calculate a loss value as an interpolation of the cross entropy between the predictions and the true labels, and the cross entropy between the predictions and the uniform distribution, where these two factors are weighed based on a value that is a normalization of the MC-based estimate of the variance. Even though their loss calculation uses the predictions of a single execution, it also requires K executions for estimating the variance. As such, the authors use five MC samples and, subsequently, five propagations of the input through the entire model during training. Instead, we use both the expectation and the variance of the outputs in our loss calculation, as propagated through the entire network in closed form approximation, constituting a more deterministic and elegant solution. Given the long-standing criticisms of MC dropout on whether its assumptions and approximations truly constitute a Bayesian method,[72–74] and the fact that sample-free Bayesian methods have outperformed MC-dropout before,[53,54] we did not consider a direct comparison with this method necessary.

Should we focus on the easy data then?

The underlying philosophy of our uncertainty-aware smoothing method is that high predictive uncertainty implies a training data sample that is, for whatever reason (e.g., difficulty, subjectivity, scarcity), difficult to model, and, as such, that our BNN should not over-penalize itself trying to memorize it. Similar assumptions have been made by past studies that focus on aleatory uncertainty,[38] and soft labels due to rater disagreement,[75] or label smoothing.[33,71] That being said, there has also been an alternate way of thinking, such as data samples that are *too easy* to model should be the ones either ignored or down-weighed, such that we avoid a flood of common samples dominating the loss calculation. A method that follows this paradigm is the focal loss,[76] of which newer versions are also heteroscedastic, i.e., dependent on the input, as the degree of focus is itself dependent on an auxiliary output of the model.[77] This is similar to our approach, albeit we are not using a separate output "head" but leverage the Bayesian predictive uncertainty. A combination of these two philosophies, and a means by which we can learn the degrees to which we should downweigh both the easy as well as the difficult samples side by side is something we would like to focus on in a future extension of this study, potentially by incorporating uncertainty decomposition methods.[38]

Generality of method

Although, in the study performed in Wu et al.,[53] the authors validated their moment-propagating BNNs on small scale, tabular datasets, in Schmitt and Roth[54] such models have also been applied on standard image classification datasets such as MNIST,[78] CIFAR,[59] and ImageNet.[79] Dera et al.[55] have gone further to image segmentation on both radar sensor and medical magnetic resonance images. Finally, Haußmann et al.[52] have used the sample-free output uncertainty in a downstream active learning framework for budgeted image classification labeling. We not only build upon such models methodologically with our adaptive label smoothing but we also apply them to a new domain, that of bioacoustic animal detection. Given the above, we see it as highly likely that the performance of our method can be transferrable to any data domain in which it is beneficial to model uncertainty, including speech and textual language processing, multimodal domains such as video, as well as graph data.

Limitations of method

Even though the parameter space required for the sample-free Bayesian models is almost equivalent to the baseline (just one additional parameter per trainable layer, as described in the supplemental experimental procedures for variance parameterization), the prediction and training times are longer (see section "execution times"). Furthermore, the activation space is double compared to the baseline as we are propagating the variances as well as the expectations. That being said, this is a known and accepted behavior in the sample-free Bayesian DL literature.[52–55] This is also reasonable, since other Bayesian considerations also require an increase in resources: e.g., MC sampling-based methods perform a number of propagations through the network that is equal to the number of MC samples, something that also introduces stochasticity in training.[54]

Conclusions and future work

Although the predictive uncertainty signal calculated by BNNs is often used to make decisions in a downstream task, such as identifying samples to annotate in active learning or addressing risk in reinforcement learning, in this article, we have used it to guide learning in the self-same task the neural network is being trained on. To that end, we have focused on deterministic (i.e., non-MC-based) BNNs that propagate feature variances along with expectations and utilized the end-to-end propagated output uncertainty to inform the degree of label smoothing that is applied in a data-specific manner. Our proposed sample-free variational Bayesian SE ResNet yields in most cases an improvement over the point-estimate baseline. Furthermore, our recommended variant with uncertainty-aware label smoothing brings further improvement in cases in which the maximum operation is used for local pooling.

Our methodology has been evaluated on two animal call-detection bioacoustics datasets, one of them introduced here for the first time, as well as in two variations pertaining to local hidden unit pooling. We find that the choice of pooling affects performance depending on the dataset, and it affects the success of uncertainty-aware label smoothing. As such, we submit that the use of uncertainty-aware label smoothing is a promising method that should be considered as a hyperparameter, to be incorporated based on validation performance. By using it, one incorporates the uncertainty value that is available to sample-free BNNs in the loss value calculation.

This work both advances work on moment-propagating BNNs that are of great use in the domain of DL and is of special interest to the application field of bioacoustics, where low signal-to-noise-ratio data often also receive weak annotation, leading to a need for soft, modest predictions that are highly calibrated (noted so far to be missing).[3,61,62] Well-calibrated model outputs with meaningful prediction probabilities are required for downstream processing either by automatic decision-making software or human experts, especially in a collaborative human-machine setting, such as active learning. Although other types of BNN are known to perform well in terms of calibration,[40,80] we have shown here that this also holds for the moment-propagating variety, with and without the use of our intelligent label smoothing.

It is important to note that this study has not been an extended comparative study of neural network architectures for acoustics as in Rizos et al.[10] Many promising point-estimate DL architectures exist, potentially focused on other data domains, that could prove to be excellent performers on one (e.g., see the experiment with the WideResNet-based BirdNET in section "SE-ResNet is a competitive point-estimate baseline," as well as Table S1) or even both the datasets we considered. Our results indicate that a sample-free Bayesian treatment of any existing point-estimate architecture is highly likely to bring further improvement, with or without our proposed uncertainty-aware label-smoothing approach. We further believe this study can stimulate research in uncertainty-aware local pooling and attention methods, in identifying informative data samples[47] in an integrated manner with focal loss,[77] and in trustworthy decision making in bioacoustics. Finally, we believe it is of interest to approach the newly introduced SAFE-MSMT dataset via a few-shot learning framework,[81] to extract as much information as possible from the limited size labeling.

EXPERIMENTAL PROCEDURES

Resource availability

Lead contact

Further information regarding the computational methodology and use of codebase should be directed to and will be fulfilled by the lead contact, G.R. (georgios.rizos12@imperial.ac.uk). Information regarding the SAFE-MSMT dataset should be addressed to R.E. (r.ewers@imperial.ac.uk), and regarding OSA-SMW to J.L. (j.lawson17@imperial.ac.uk).

Materials availability

This study did not generate new unique materials or reagents.

Data and code availability

The latest version of the code can be found at https://github.com/glam-imperial/sample-free-uncertainty-label-smoothing under DOI through Zenodo: https://doi.org/10.5281/zenodo.10253149 [82] and is publicly available as of the date of publication. The SAFE-MSMT dataset introduced in this paper is to be found at Zenodo: https://doi.org/10.5281/zenodo.7740620 [83]; the contact for this dataset is R.E. The OSA-SMW data reported in this paper will be shared by J.L. upon request.

Description of multi-attentive SE-ResNet

The base DL architecture we use in this study is a close variant of the best performing method from the comparative study in Rizos et al.[10] The sample-free Bayesian treatment is applied on the same architecture, whether uncertainty-aware label smoothing is used or not. Table 1 summarizes the number of layers used and related parameters.

We can divide the architecture in three modules: (1) the core, audio processing module, which produces a sequence of learnt audio embeddings and is based on convolutional layers, residual blocks, local pooling (maximum or attentive), and SE blocks; (2) the multiple-head, attention mechanism for weighted average pooling of the embeddings; and (3) the top module, a set of dense layers that process the averaged, recording-wide neural representation, where each layer makes a prediction corresponding to a separate binary call-detection task. There is one such layer for the OSA-SMW and 30 for the SAFE-MSMT dataset. We extract spectrograms from sound waveform sampled at a rate of 16 kHz, by using a fast Fourier transform window of 128 ms, sliding at a hop length of 10 ms. Given a 3-s clip, we extract 128 Mel coefficients and end up with a log-Mel spectrogram with sequence length equal to 300.

As seen in Table 1, the log-Mel spectrogram is first processed by a block (ConvBlock) of two convolutional layers, each with 64 filters and ReLU activations, and followed by a pooling operation without padding. The pooling operation can be either max- or attentive pooling. Then, the hidden units are processed by four blocks (SEBlock), where each is composed of two residual layers with SE mechanisms, and is followed by a pooling operation. The core module concludes with another ConvBlock, where the convolutions learn 1,024 filters, but this time not followed by pooling. In all cases, the convolutional layers learn 3×3 filters and corresponding biases, and the pooling operations are subsampling at a 2×2 ratio.

The above module transforms a log-Mel spectrogram input into a hidden tensor with sequence length of nine, width of four, and 1,024 features. We want to perform global pooling across the sequence length, and so first reshape the tensor to $(9, 4096)$. We then learn four weighted sequence-averaging operations, using four attention heads. Each head corresponds to a learnt linear transformation of each embedding frame to a single energy value, and the calculation of a probability vector by passing the energy values from the sequence through a softmax function. These probabilities are used for weighted averaging, leading to an averaged embedding per attention head; those are then concatenated to provide a single, sequence-wide representation of the input audio clip. This is processed by the top module, where the dense layer that corresponds to each task avails of the common base model for shared feature extraction. Each dense layer produces one logit per data sample, which is passed through a sigmoid function such that we obtain the probability that the sample is positive.

Epistemic uncertainty-aware label smoothing

We need to quantify the belief that an input sample has been noisily annotated, and as such the prediction error for it should contribute less to the loss value calculation. We design such a measure by adhering to the following desiderata: (1) it is in the $[0, 1]$ range, such that it can serve as the label-smoothing probability; (2) it is positively correlated to the propagated, predictive variance in order to reflect BNN uncertainty about the input sample categorization; and (3) it is also positively correlated to overconfident (i.e., close to 1) predictions, such that moderate predictions do not receive feedback reinforcement.

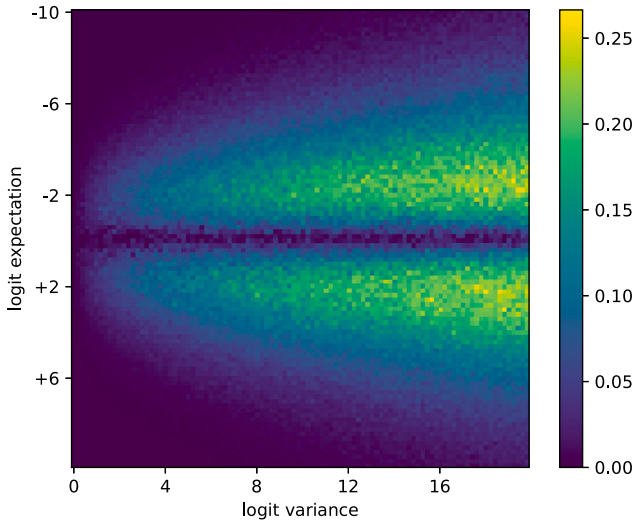


Figure 2. The value of our proposed adaptive, uncertainty-aware smoothing probability given the expectation and variance of the logit

For close to 0 logit uncertainties, the smoothing probability α_i^t is also close to 0. For higher logit uncertainties $\mathbb{V}[h_{i,L}^t]$, α_i^t is higher for predictions that are closer to the extreme values of either 0 or 1. For moderate predictions close to 0.5, α_i^t is closer to 0, thus encouraging learning from the true signal instead of reinforcing a moderate prediction behavior.

Consider the expected logit output $\mathbb{E}[h_{i,L}^t]$ of a dense prediction layer for the i-th acoustic data sample, where L denotes the last layer index and t denotes the task corresponding to that prediction layer. If we *do* utilize the logit variance $\mathbb{V}[h_{i,L}^t]$ and transform the normally distributed random variable via a sigmoid function (as detailed in the supplemental experimental procedures, section "sample-free variational attentive SE-ResNet"), we get the fully propagated, Bayesian expectation and variance of $y_{i,\mathrm{POS}}^{\mathrm{t,Bayes}}$, i.e., the probability that the input sample is from the positive (POS) class. Inversely, if we opt for a maximum *a posteriori* (MAP) approach for that final layer, by *not* utilizing the logit variance, we transform the logit expectation via the sigmoid and denote the probability by $y_{i,\mathrm{POS}}^{\mathrm{t,MAP}}$.

$y_{i,\mathrm{POS}}^{\mathrm{t,MAP}}$ would still benefit from the moment propagation up until the final layer in terms of the learnt features and logits h_l (for l up to but excluding L), as well as from the Bayesian regularization for all layers. However, final-layer MAP makes the information encoded in the propagated uncertainty unavailable in the calculation of the predictive probability distribution. Inversely, $y_{i,\mathrm{POS}}^{\mathrm{t,Bayes}}$ gets the full benefits of the Bayesian approach. A fully Bayesian treatment of even just the final layer has been shown to have a positive benefit on addressing overconfidence, even when the rest of the model is parameterized with point-estimate weights.[84]

We, thus, attempt to capture this additional, Bayesian uncertainty information by defining the data-sample-specific smoothing probability as

$$\alpha_i^t = \left| y_{i,\mathrm{POS}}^{\mathrm{t,MAP}} - y_{i,\mathrm{POS}}^{\mathrm{t,Bayes}} \right|. \qquad \text{(Equation 2)}$$

For a binary call-detection task, this is equivalent to the Manhattan distance between the corresponding two-element discrete predictive probability distributions multiplied by two. A visualization of our adaptive smoothing probability given ranges of logit expectations and variances can be found in Figure 2.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.patter.2024.100932.

ACKNOWLEDGMENTS

G.R. would like to acknowledge the Engineering and Physical Sciences Research Council (EPSRC) grant no. 2021037.

AUTHOR CONTRIBUTIONS

G.R. conceived and designed the proposed methodology, coded the moment-propagating BNNs and smoothing methods, executed the experiments, wrote the article, and prepared figures. J.L. collected and annotated the OSA-SMW dataset and contributed in designing the predictive task and writing related parts. S.M. annotated the SAFE-MSMT dataset. P.S. contributed in coding the dense Bayesian neural layer and the automatic relevance determination prior and proposed the use of cold posteriors for training. X.W. preprocessed the SAFE-MSMT dataset, prepared initial versions of related figures and descriptions of related parts, and executed exploratory experiments on SAFE-MSMT. C.B.-L., R.E., and B.W.S. supervised the research. All authors discussed the results and commented on/edited the manuscript.

DECLARATION OF INTERESTS

G.R. is affiliated with the University of Cambridge. This work was performed during his PhD candidacy at Imperial College London. J.L. is also affiliated with the UK Centre for Ecology and Hydrology. P.S. is now affiliated with Advai Ltd. P.S. and X.W. worked on this study as MSc students at Imperial College London. B.W.S. is also affiliated with the Technical University of Munich and audEERING GmbH.

Received: October 7, 2022
Revised: March 1, 2023
Accepted: January 19, 2024
Published: February 12, 2024

REFERENCES

1. Witmer, G.W. (2005). Wildlife population monitoring: some practical considerations. Wildl. Res. *32*, 259–263. https://doi.org/10.1071/WR04003.
2. Tuia, D., Kellenberger, B., Beery, S., Costelloe, B.R., Zuffi, S., Risse, B., Mathis, A., Mathis, M.W., van Langevelde, F., Burghardt, T., et al. (2022). Perspectives in machine learning for wildlife conservation. Nat. Commun. *13*, 1–15. https://doi.org/10.1038/s41467-022-27980-y.
3. Stowell, D. (2022). Computational bioacoustics with deep learning: a review and roadmap. PeerJ *10*, e13152. https://doi.org/10.7717/peerj.13152.
4. Turner, W. (2014). Sensing biodiversity. Science *346*, 301–302. https://doi.org/10.1126/science.1256014.
5. Veit, A., Alldrin, N., Chechik, G., Krasin, I., Gupta, A., and Belongie, S. (2017). Learning from noisy large-scale datasets with minimal supervision. In Proceedings of the Conference on Computer Vision and Pattern Recognition (IEEE/CVF), pp. 839–847. https://doi.org/10.1109/CVPR.2017.696.
6. Arroyo-Rodríguez, V., and Fahrig, L. (2014). Why is a landscape perspective important in studies of primates? Am. J. Primatol. *76*, 901–909. https://doi.org/10.1002/ajp.22282.
7. Hershey, S., Chaudhuri, S., Ellis, D.P.W., Gemmeke, J.F., Jansen, A., Moore, R.C., Plakal, M., Platt, D., Saurous, R.A., Seybold, B., et al. (2017). CNN architectures for large-scale audio classification. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing (IEEE), pp. 131–135. https://doi.org/10.1109/ICASSP.2017.7952132.
8. He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the Conference on Computer Vision and Pattern Recognition (IEEE/CVF), pp. 770–778. https://doi.org/10.1109/CVPR.2016.90.
9. Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., and Plumbley, M.D. (2020). PANNs: Large-scale pretrained audio neural networks for audio pattern recognition. IEEE/ACM Trans. Audio Speech Lang. Process. *28*, 2880–2894. https://doi.org/10.1109/TASLP.2020.3030497.
10. Rizos, G., Lawson, J., Han, Z., Butler, D., Rosindell, J., Mikolajczyk, K., Banks-Leite, C., and Schuller, B.W. (2021). Multi-attentive detection of the spider monkey qhinny in the (actual) wild. Proceedings of Interspeech (ISCA), 471–475. https://doi.org/10.21437/Interspeech.2021-1969.
11. Hong, S., Zou, Y., and Wang, W. (2020). Gated multi-head attention pooling for weakly labelled audio tagging. Proceedings of Interspeech (ISCA), 816–820. https://doi.org/10.21437/Interspeech.2020-1197.
12. Naranjo-Alcazar, J., Perez-Castanos, S., Zuccarello, P., and Cobos, M. (2020). Acoustic scene classification with squeeze-excitation residual networks. IEEE Access *8*, 112287–112296. https://doi.org/10.1109/ACCESS.2020.3002761.
13. Kiskin, I., Cobb, A.D., Sinka, M., Willis, K., and Roberts, S.J. (2021). Automatic acoustic mosquito tagging with bayesian neural networks. In Proceedings of Joint European Conference on Machine Learning and Knowledge Discovery in Databases (Springer), pp. 351–366. https://doi.org/10.1007/978-3-030-86514-6_22.
14. Dufourq, E., Durbach, I., Hansford, J.P., Hoepfner, A., Ma, H., Bryant, J.V., Stender, C.S., Li, W., Liu, Z., Chen, Q., et al. (2021). Automated detection of hainan gibbon calls for passive acoustic monitoring. Remote Sens. Ecol. Conserv. *7*, 475–487. https://doi.org/10.1002/rse2.201.
15. Tzirakis, P., Shiarella, A., Ewers, R., and Schuller, B.W. (2020). Computer audition for continuous rainforest occupancy monitoring: The case of bornean gibbons' call detection. Proceedings of Interspeech (ISCA), 1211–1215. https://doi.org/10.21437/Interspeech.2020-2655.
16. Bahdanau, D., Cho, K.H., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In Proceedings of the International Conference on Learning Representations (Preprint at arXiv). https://doi.org/10.48550/arXiv.1409.0473.
17. Luong, M.T., Pham, H., and Manning, C.D. (2015). Effective approaches to attention-based neural machine translation. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1412–1421. https://doi.org/10.18653/v1/D15-1166.
18. Hu, J., Shen, L., and Sun, G. (2018). Squeeze-and-excitation networks. In Proceedings of the Conference on Computer Vision and Pattern Recognition (IEEE/CFV), pp. 7132–7141. https://doi.org/10.1109/CVPR.2018.00745.
19. Zhang, Z., Wu, B., and Schuller, B.W. (2019). Attention-augmented end-to-end multi-task learning for emotion prediction from speech. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing (IEEE), pp. 6705–6709. https://doi.org/10.1109/ICASSP.2019.8682896.
20. Kahl, S., Wood, C.M., Eibl, M., and Klinck, H. (2021). Birdnet: A deep learning solution for avian diversity monitoring. Ecol. Inf. *61*, 101236. https://doi.org/10.1016/j.ecoinf.2021.101236.
21. Ruan, W., Wu, K., Chen, Q., and Zhang, C. (2022). Resnet-based bio-acoustics presence detection technology of hainan gibbon calls. Appl. Acoust. *198*, 108939. https://doi.org/10.1016/j.apacoust.2022.108939.
22. Zagoruyko, S., and Komodakis, N. (2016). Wide residual networks. In Procedings of the British Machine Vision Conference (British Machine Vision Association).
23. Pérez-Granados, C. (2023). Birdnet: applications, performance, pitfalls and future opportunities. Ibis *165*, 1068–1075. https://doi.org/10.1111/ibi.13193.
24. Shiu, Y., Palmer, K.J., Roch, M.A., Fleishman, E., Liu, X., Nosal, E.M., Helble, T., Cholewiak, D., Gillespie, D., and Klinck, H. (2020). Deep neural networks for automated detection of marine mammal species. Sci. Rep. *10*, 607–612. https://doi.org/10.1038/s41598-020-57549-y.
25. Hantke, S., Cummins, N., and Schuller, B.W. (2018). What is my dog trying to tell me? The automatic recognition of the context and perceived emotion of dog barks. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing (IEEE), pp. 5134–5138. https://doi.org/10.1109/ICASSP.2018.8461757.
26. Oikarinen, T., Srinivasan, K., Meisner, O., Hyman, J.B., Parmar, S., Fanucci-Kiss, A., Desimone, R., Landman, R., and Feng, G. (2019).

Deep convolutional network for animal sound classification and source attribution using dual audio recordings. J. Acoust. Soc. Am. *145*, 654–662. https://doi.org/10.1121/1.5087827.

27. Clink, D.J., and Klinck, H. (2019). Gibbonfindr: An R package for the detection and classification of acoustic signals. Preprint at arXiv. https://doi.org/10.48550/arXiv.1906.02572.
28. Goëau, H., Glotin, H., Vellinga, W.P., Planqué, R., and Joly, A. (2016). Lifeclef bird identification task 2016: The arrival of deep learning. In Proceedings of CLEF: Conference and Labs of the Evaluation Forum, pp. 440–449.
29. Rovithis, E., Moustakas, N., Vogklis, K., Drossos, K., and Floros, A. (2021). Towards citizen science for smart cities: A framework for a collaborative game of bird call recognition based on internet of sound practices. Preprint at arXiv. https://doi.org/10.48550/arXiv.2103.16988.
30. Guo, C., Pleiss, G., Sun, Y., and Weinberger, K.Q. (2017). On calibration of modern neural networks. In Proceedings of the International Conference on Machine Learning (PMLR), pp. 1321–1330.
31. Tomsett, R., Preece, A., Braines, D., Cerutti, F., Chakraborty, S., Srivastava, M., Pearson, G., and Kaplan, L. (2020). Rapid trust calibration through interpretable and uncertainty-aware ai. Patterns (N. Y). *1*, 100049. https://doi.org/10.1016/j.patter.2020.100049.
32. Tomani, C., and Buettner, F. (2021). Towards trustworthy predictions from deep neural networks with fast adversarial calibration. Proc. AAAI Conf. Artif. Intell. *35*, 9886–9896.
33. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR/CVF), pp. 2818–2826. https://doi.org/10.1109/CVPR.2016.308.
34. Singh, A., Bay, A., Sengupta, B., and Mirabile, A. (2021). On the dark side of calibration for modern neural networks. In ICML Workshop on Uncertainty and Robustness in Deep Learning.
35. Lukasik, M., Bhojanapalli, S., Menon, A., and Kumar, S. (2020). Does label smoothing mitigate label noise? Proceedings of the International Conference on Machine Learning (PMLR), 6448–6458.
36. Wei, J., Liu, H., Liu, T., Niu, G., Sugiyama, M., and Liu, Y. (2021). To smooth or not? when label smoothing meets noisy labels. In Proceedings of the International Conference on Machine Learning (PMLR), pp. 23589–23614.
37. Wang, D.B., Feng, L., and Zhang, M.L. (2021). Rethinking calibration of deep neural networks: Do not be afraid of overconfidence. In Proceedings of Advances in Neural Information Processing Systems, pp. 11809–11820.
38. Kendall, A., and Gal, Y. (2017). What uncertainties do we need in bayesian deep learning for computer vision? In Proceedings of Advances in Neural Information Processing Systems, pp. 5580–5590.
39. Mackay, D.J.C. (1992). Bayesian methods for adaptive models. PhD Thesis (California Institute of Technology).
40. Maddox, W.J., Garipov, T., Izmailov, P., Vetrov, D., and Wilson, A.G. (2019). A simple baseline for bayesian uncertainty in deep learning. In Proceedings of Advances in Neural Information Processing Systems, pp. 13153–13164.
41. Wilson, A.G. (2020). The case for bayesian deep learning. Preprint at arXiv. https://doi.org/10.48550/arXiv.2001.10995.
42. Wang, H., and Yeung, D.Y. (2020). A survey on bayesian deep learning. ACM Comput. Surv. *53*, 1–37. https://doi.org/10.1145/3409383.
43. Blei, D.M., Kucukelbir, A., and McAuliffe, J.D. (2017). Variational inference: A review for statisticians. J. Am. Stat. Assoc. *112*, 859–877. https://doi.org/10.1080/01621459.2017.1285773.
44. Zhang, C., Bütepage, J., Kjellström, H., and Mandt, S. (2019). Advances in variational inference. IEEE Trans. Pattern Anal. Mach. Intell. *41*, 2008–2026. https://doi.org/10.1109/TPAMI.2018.2889774.
45. Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. (2015). Weight uncertainty in neural networks. In Proceedings of the International conference on machine learning (PMLR), pp. 1613–1622.
46. Gal, Y., and Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In Proceedings of the International conference on machine learning (PMLR), pp. 1050–1059.
47. Rizos, G., and Schuller, B.W. (2019). Modelling sample informativeness for deep affective computing. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing (IEEE), pp. 3482–3486. https://doi.org/10.1109/ICASSP.2019.8683729.
48. Wang, S., and Manning, C. (2013). Fast dropout training. In Proceedings of the International Conference on Machine Learning (PMLR), pp. 118–126.
49. Kingma, D.P., Salimans, T., and Welling, M. (2015). Variational dropout and the local reparameterization trick. In Proceedings of Advances in Neural Information Processing Systems, pp. 2575–2583.
50. Roth, W., and Pernkopf, F. (2016). Variational inference in neural networks using an approximate closed-form objective. In Proceedings of the NIPS Workshop on Bayesian Deep Learning.
51. Shridhar, K., Laumann, F., and Liwicki, M. (2018). Uncertainty estimations by softplus normalization in bayesian convolutional neural networks with variational inference. Preprint at arXiv. https://doi.org/10.48550/arXiv.1806.05978.
52. Haußmann, M., Hamprecht, F., and Kandemir, M. (2019). Deep active learning with adaptive acquisition. In Proceedings of the International Joint Conference on Artificial Intelligence (ACM), pp. 2470–2476.
53. Wu, A., Nowozin, S., Meeds, E., Turner, R.E., Hernandez-Lobato, J.M., and Gaunt, A.L. (2018). Deterministic variational inference for robust bayesian neural networks. In Proceedings of the International Conference on Learning Representations (Preprint at arXiv). https://doi.org/10.48550/arXiv.1810.03958.
54. Schmitt, J., and Roth, S. (2021). Sampling-free variational inference for neural networks with multiplicative activation noise. In Proceedings of DAGM German Conference on Pattern Recognition (Springer), pp. 33–47. https://doi.org/10.1007/978-3-030-92659-5_3.
55. Dera, D., Bouaynaya, N.C., Rasool, G., Shterenberg, R., and Fathallah-Shaykh, H.M. (2021). Premium-CNN: Propagating uncertainty towards robust convolutional neural networks. IEEE Trans. Signal Process. *69*, 4669–4684. https://doi.org/10.1109/TSP.2021.3096804.
56. Goodfellow, I.J., Shlens, J., and Szegedy, C. (2014). Explaining and harnessing adversarial examples. Preprint at arXiv. https://doi.org/10.48550/arXiv.1412.6572.
57. Wang, P., Bouaynaya, N.C., Mihaylova, L., Wang, J., Zhang, Q., and He, R. (2020). Bayesian neural networks uncertainty quantification with cubature rules. In Proceedings of the International Joint Conference on Neural Networks (IEEE), pp. 1–7. https://doi.org/10.1109/IJCNN48605.2020.9207214.
58. Carannante, G., Bouaynaya, N.C., and Mihaylova, L. (2021). An enhanced particle filter for uncertainty quantification in neural networks. In Proceedings of the International Conference on Information Fusion (IEEE), pp. 1–7. https://doi.org/10.23919/FUSION49465.2021.9626883.
59. Krizhevsky, A. (2009). Learning multiple layers of features from tiny images. Master's Thesis (University of Toronto).
60. Tzelepis, C., and Patras, I. (2021). Uncertainty propagation in convolutional neural networks: Technical report. Preprint at arXiv. https://doi.org/10.48550/arXiv.2102.06064.
61. Stowell, D., Wood, M.D., Pamuła, H., Stylianou, Y., and Glotin, H. (2019). Automatic acoustic detection of birds through deep learning: the first bird audio detection challenge. Methods Ecol. Evol. *10*, 368–380. https://doi.org/10.1111/2041-210x.13103.
62. Kitzes, J., and Schricker, L. (2019). The necessity, promise and challenge of automated biodiversity surveys. Environ. Conserv. *46*, 247–250. https://doi.org/10.1017/S0376892919000146.
63. Krokos, V., Bui Xuan, V., Bordas, S.P.A., Young, P., and Kerfriden, P. (2022). A bayesian multiscale cnn framework to predict local stress fields in structures with microscale features. Comput. Mech. *69*, 733–766. https://doi.org/10.1007/s00466-021-02112-3.

64. Haußmann, M., Hamprecht, F.A., and Kandemir, M. (2020). Sampling-free variational inference of bayesian neural networks by variance backpropagation. In Proceedings of Uncertainty in Artificial Intelligence (PMLR), pp. 563–573. https://doi.org/10.48550/arXiv.1805.07654.

65. Gao, Z., Wang, L., and Wu, G. (2019). Lip: Local importance-based pooling. In Proceedings of the International Conference on Computer Vision (IEEE), pp. 3355–3364. https://doi.org/10.1007/s11263-022-01707-4.

66. Stergiou, A., and Poppe, R. (2023). Adapool: Exponential adaptive pooling for information-retaining downsampling. IEEE Trans. Image Process. *32*, 251–266. https://doi.org/10.1109/TIP.2022.3227503.

67. Ewers, R.M., Didham, R.K., Fahrig, L., Ferraz, G., Hector, A., Holt, R.D., Kapos, V., Reynolds, G., Sinun, W., Snaddon, J.L., and Turner, E.C. (2011). A large-scale forest fragmentation experiment: the stability of altered forest ecosystems project. Philos. Trans. R. Soc. Lond. B Biol. Sci. *366*, 3292–3302. https://doi.org/10.1098/rstb.2011.0049.

68. Gal, Y., Islam, R., and Ghahramani, Z. (2017). Deep bayesian active learning with image data. In Proceedings of International Conference on Machine Learning (PMLR)), pp. 1183–1192.

69. Depeweg, S., Hernandez-Lobato, J.M., Doshi-Velez, F., and Udluft, S. (2018). Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning. In Proceedings of International Conference on Machine Learning (PMLR), pp. 1184–1193.

70. Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network. Preprint at arXiv. https://doi.org/10.48550/arXiv.1503.02531.

71. Seo, S., Seo, P.H., and Han, B. (2019). Learning for single-shot confidence calibration in deep neural networks through stochastic inferences. In Proceedings of the Conference on Computer Vision and Pattern Recognition (IEEE/CVF)), pp. 9030–9038. https://doi.org/10.1109/CVPR.2019.00924.

72. Osband, I. (2016). Risk versus uncertainty in deep learning: Bayes, bootstrap and the dangers of dropout. In NIPS workshop on bayesian deep learning.

73. Verdoja, F., and Kyrki, V. (2020). Notes on the behavior of MC dropout. Preprint at arXiv. https://doi.org/10.48550/arXiv.2008.02627.

74. Folgoc, L.L., Baltatzis, V., Desai, S., Devaraj, A., Ellis, S., Martinez Manzanera, O.E., Nair, A., Qiu, H., Schnabel, J., and Glocker, B. (2021). Is MC dropout bayesian?. Preprint at arXiv. https://doi.org/10.48550/arXiv.2110.04286.

75. Chou, H.C., and Lee, C.C. (2019). Every rating matters: Joint learning of subjective labels and individual annotators for speech emotion classification. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing (IEEE)), pp. 5886–5890. https://doi.org/10.1109/ICASSP.2019.8682170.

76. Lin, T.Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017). Focal loss for dense object detection. In Proceedings of the International Conference on Computer Vision (IEEE), pp. 2980–2988. https://doi.org/10.1109/ICCV.2017.324.

77. Li, X., Wang, W., Hu, X., Li, J., Tang, J., and Yang, J. (2021). Generalized focal loss v2: Learning reliable localization quality estimation for dense object detection. In Proceedings of the Conference on Computer Vision and Pattern Recognition (IEEE/CVF)), pp. 11632–11641. https://doi.org/10.1109/CVPR46437.2021.01146.

78. LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. Proc. IEEE *86*, 2278–2324. https://doi.org/10.1109/5.726791.

79. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., and Li, F.F. (2009). Imagenet: A large-scale hierarchical image database. In Proceedings of the Conference on Computer Vision and Pattern Recognition (IEEE), pp. 248–255. https://doi.org/10.1109/CVPR.2009.5206848.

80. Osawa, K., Swaroop, S., Khan, M.E.E., Jain, A., Eschenhagen, R., Turner, R.E., and Yokota, R. (2019). Practical deep learning with bayesian principles. In Roceedings of Advances in Neural Information Processing Systems (ACM), pp. 4287–4299.

81. Nolasco, I., Singh, S., Vidaña-Vila, E., Grout, E., Morford, J., Emmerson, M.G., Jensen, F.H., Kiskin, I., Whitehead, H., Strandburg-Peshkin, A., et al. (2022). Few-shot bioacoustic event detection at the dcase 2022 challenge. In Detection and Classification of Acoustic Scenes and Events.

82. Rizos, G. (2023). Code for the article "Propagating Variational Model Uncertainty for Bioacoustic Call Label Smoothing". Zenodo. https://doi.org/10.5281/zenodo.10253149.

83. Trigg, L., Mitchell, S., and Ewers, R.M. (2023). Assessment of acoustic indices for monitoring phylogenetic and temporal patterns of biodiversity in tropical forests. Zenodo. https://doi.org/10.5281/zenodo.7740620.

84. Kristiadi, A., Hein, M., and Hennig, P. (2020). Being bayesian, even just a bit, fixes overconfidence in relu networks. In Proceedings of the International conference on machine learning (PMLR), pp. 5436–5446.